

## 5.- Selección de las distribuciones de entrada

### 5.1- Elección de la familia de distribuciones

El primer paso en la selección de la distribución de entrada es decidir qué familia de distribuciones (exponencial, gamma, normal, Poisson, etc...) parece ser más apropiada, sin preocuparse, de momento, de los valores en concreto que deben tomar los parámetros de la distribución. En esta sección se describen métodos generales que pueden servir de ayuda a la hora de escoger qué familia de distribuciones usar para generar una variable aleatoria particular en una simulación.

#### 5.1.1- En base al conocimiento teórico

En algunas ocasiones se dispone a priori del conocimiento suficiente acerca de la variable aleatoria como para poder seleccionar la distribución o al menos para poder descartar algunas otras. Esto se hace sobre la base de conocimientos teóricos y no es necesaria ninguna medida experimental para ello. Para ello es conveniente tener en mente las características de las distribuciones que hemos visto:

*Distribución uniforme  $U(a,b)$ :* es equiprobable que  $X$  tome cualquier valor en el intervalo  $[a,b]$ .

*Distribución exponencial:* suele ser un modelo apropiado para intervalos entre eventos de interés o tiempos de espera hasta que ocurra un evento.

*Distribución gamma:* suele emplearse para modelar el tiempo necesario para completar una tarea.

*Distribución normal:* el teorema del límite central induce a suponer que una distribución normal será apropiada cuando las observaciones se construyan a partir de sumas o promedios de variables aleatorias independientes entre si y tales que cada una de ellas realiza una pequeña contribución a la suma.

*Distribución de Bernoulli:* si la variable aleatoria representa el resultado de un experimento con dos únicos posibles resultados (experimento de Bernoulli).

*Distribución uniforme discreta:* si la variable aleatoria es el resultado de un experimento con varios posibles resultados, todos ellos igualmente probables.

*Distribución binomial:* surge cuando estamos interesados en el número de miembros de un grupo de individuos que poseen cierta característica. El grupo de individuos ha sido escogido aleatoriamente y posee un tamaño fijo y conocido a priori. Cada individuo tiene la misma probabilidad de poseer la característica. El hecho de que un individuo posea la característica es independiente del hecho de que otro la posea. El hecho de saber que ciertos individuos posean la característica no nos ayuda a decidir si otro individuo en concreto la tiene.

*Distribución geométrica:* surge cuando estamos interesados en conocer el número de repeticiones de un experimento de Bernoulli que deben hacerse hasta que se da uno determinado de los dos posibles resultados.

*Distribución negativa binomial:* surge cuando estamos interesados en el número de experimentos de Bernoulli "fallidos" antes de que se de un determinado número de veces uno determinado de los dos posibles resultados.

*Distribución de Poisson:* surge normalmente cuando estamos interesados en el número de eventos aleatorios que suceden en un intervalo fijo.

Veamos dos métodos gráficos para seleccionar el tipo de distribución apropiado: *histogramas y gráficas de probabilidad*.

### 5.1.2- Histogramas

Un histograma es básicamente una estimación gráfica de la función densidad de probabilidad correspondiente a la distribución de nuestros datos,  $X_1, \dots, X_n$ . Las funciones densidad de probabilidad de las distribuciones teóricas en algunos casos tienen formas reconocibles, con lo cual, el histograma puede constituir una ayuda para saber a que familias de distribuciones debe hacerse el ajuste.

En el caso continuo, para construir un histograma debe dividirse el rango de valores cubierto por los datos experimentales en  $k$  intervalos disjuntos  $[b_0, b_1), [b_1, b_2), \dots, [b_{k-1}, b_k)$ , de igual longitud,  $\Delta b = b_j - b_{j-1}$ . Para  $j=1, 2, \dots, k$ , sea  $q_j$  la proporción de los datos experimentales,  $X_1, \dots, X_n$ , que se encuentran en el intervalo  $[b_{j-1}, b_j)$ . Finalmente, se define la función, constante a tramos:

$$h(x) = \begin{cases} 0 & \text{si } x < b_0 \\ q_j & \text{si } b_{j-1} \leq x < b_j \text{ para } j = 1, \dots, k \\ 0 & \text{si } b_k \leq x \end{cases}$$

La forma de la función  $h(x)$  puede compararse con la forma (no con la escala o posición) de la función densidad de probabilidad de las distribuciones teóricas estándar, para saber qué familia de distribuciones interesa ajustar a los datos experimentales.

En el caso discreto, la construcción del histograma (o diagrama de barras) es más sencilla ya que no es preciso definir ni los intervalos ni agrupar los datos.

### 5.1.3- Gráficas de probabilidad (aplicable sólo al caso continuo)

La distribución de probabilidad acumulada de una variable  $X$  aleatoria se define:

$$F(x) = P\{X \leq x\}$$

con lo cual, si  $X$  tiene la misma distribución que los datos  $X_i$ , una aproximación razonable a  $F(x)$  es entonces la proporción de los datos  $X_i$  que son menores o iguales que  $x$ . Cabría pues, comparar la distribución de probabilidad acumulada obtenida a partir de los datos experimentales, con las distribuciones de probabilidad acumulada de las distribuciones estándar, sin embargo, estas suelen tener forma de "S", con lo cual la comparación visual no suele ser demasiado esclarecedora.

Existen técnicas para reducir el problema de la comparación de funciones distribución de probabilidad acumulada a decidir cual, de entre varias gráficas, se asemeja más a una recta. La que expondremos esta basada en la comparación de los cuantiles o puntos críticos de las distribuciones. El cuantil  $q$  ( $0 < q < 1$ ) de una distribución  $F$  es un número  $x_q$  que satisface  $F(x_q) = q$ . Si  $F^{-1}$  nota la inversa de la función de distribución  $F$ , la fórmula del cuantil  $q$  de  $F$  es:

$$x_q = F^{-1}(q)$$

Si  $F$  y  $G$  son dos funciones distribución de probabilidad acumulada, son iguales,  $F=G$ , si y sólo si cada uno de los cuantiles de  $F$  es igual al correspondiente cuantil de  $G$ . Así pues, si  $x_q$  e  $y_q$  son respectivamente los cuantiles  $q$  de  $F$  y  $G$ , la representación gráfica de los puntos  $(x_q, y_q)$  debería ser una línea recta con una pendiente de  $45^\circ$  que pase por el origen, ya que  $x_q = y_q$  para todo  $q$ .

Si las variables aleatorias correspondientes a  $F$  y  $G$  difieren sólo en la posición y la escala, entonces existen dos reales  $\gamma$  y  $\beta$  ( $\beta > 0$ ) de modo que:

$$G(x) = F\left(\frac{x - \gamma}{\beta}\right) \quad \text{para todo } x$$

En este caso, para todo  $q$  se satisface:  $y_q = \gamma + \beta x_q$ , con lo cual la representación gráfica de los puntos  $(x_q, y_q)$  será una recta de pendiente diferente de  $45^\circ$  y que no pasará por el origen.

Así pues, distribuciones con la misma forma (aunque posiblemente con diferente posición y escala) dan lugar a líneas rectas. La representación gráfica de los pares  $(x_q, y_q)$  recibe el nombre de "gráficas de probabilidad".

Las gráficas de probabilidad constituyen una herramienta útil para decidir si la distribución empírica  $\tilde{F}_n$ , definida a partir de los datos experimentales, tiene la misma forma que alguna función de distribución acumulada de las familias estándar.

Veamos cómo puede definirse la distribución  $\tilde{F}_n$  a partir de los datos experimentales. Para ello, notamos  $X_{(i)}$  al  $i$ -ésimo dato experimental más pequeño, es decir,  $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ . La distribución empírica  $\tilde{F}_n$  se define como:

$$\tilde{F}_n(X_{(i)}) = \frac{i}{n} \quad \text{para } i:1,2,\dots,n$$

ya que es la proporción de los datos,  $X_1, \dots, X_n$ , que es menor o igual que  $X_{(i)}$ . Sin embargo, definida la probabilidad acumulada de esta manera, presenta la desventaja de que vale 1 para un valor finito de  $x$ ,  $\tilde{F}_n(X_{(n)}) = 1$ . Para evitar este inconveniente (que la variable aleatoria no pueda tomar valores mayores que  $X_{(n)}$ ), normalmente se modifica ligeramente la definición de  $\tilde{F}_n$ :

$$\tilde{F}_n(X_{(i)}) = \frac{i-0.5}{n} \quad \text{para } i:1,2,\dots,n$$

Asumiendo esta última definición, puede representarse gráficamente la función  $\tilde{F}_n$  dibujando los  $n$  puntos  $\left(X_{(1)}, \frac{0.5}{n}\right), \left(X_{(2)}, \frac{1.5}{n}\right), \dots, \left(X_{(n)}, \frac{n-0.5}{n}\right)$

Una vez decidido el tipo de distribución teórica con que se quiere comparar la distribución  $\tilde{F}_n$ , debe ajustarse la distribución teórica a los datos experimentales para calcular su factor de forma (si corresponde). Más adelante explicaremos cómo puede realizarse este ajuste. Los parámetros de posición y escala, por el contrario, no es preciso calcularlos; suele escogerse el parámetro de posición igual a cero y el de escala igual a 1, aunque cualquier otra elección (de entre las admisibles) sería válida. Notemos  $F$  a la distribución estándar resultante.

Para la distribución uniforme puede escogerse la distribución de probabilidad acumulada de  $U(0,1)$ . Para la distribución exponencial, la de  $\text{expo}(1)$ . Para la distribución normal, la de  $N(0,1)$ . Para la distribución gamma se calcula el parámetro de forma,  $\alpha$ , mediante ajuste, imponiendo  $\beta = 1$ .

Queremos comparar  $F$  y  $\tilde{F}_n$ , para lo cual representaremos gráficamente los pares de cuantiles para  $q = \frac{i-0.5}{n}$ , con  $i:1,2,\dots,n$ , para lo cual habremos de tener en cuenta que el cuantil  $\frac{i-0.5}{n}$  de  $\tilde{F}_n$  es precisamente  $X_{(i)}$ . Así pues, representamos los puntos

$$\left(X_{(i)}, F^{-1}\left(\frac{i-0.5}{n}\right)\right) \quad \text{para } i:1,2,\dots,n$$

y si la representación es una recta (con independencia de su pendiente, o de si pasa o no por el origen), sabemos que, a falta de ajustar los parámetros de posición y de escala,  $F$  es una buena distribución de probabilidad acumulada para nuestros datos.

Este procedimiento es sencillo cuando la fórmula  $F^{-1}\left(\frac{i-0.5}{n}\right)$  es fácilmente calculable, como en el caso de las distribuciones uniforme y exponencial, en cambio, para las distribuciones normal, gamma y beta, entre otras, debe calcularse numéricamente.

Este método generalmente no es aplicable al caso discreto, en el cual, en general,  $F^{-1}(y)$  no está bien definido.

### 5.1.4.- Test de normalidad de los momentos

Existen varios test que permiten determinar el alejamiento respecto de la distribución normal de un conjunto de observaciones. Uno de ellos es el test de los momentos.

El tercer y cuarto momento de la distribución se definen:

$$m_3 = E\{(X - \mu)^3\}$$

$$m_4 = E\{(X - \mu)^4\}$$

A partir de un conjunto de observaciones  $X_1, \dots, X_n$  pueden estimarse de la forma:

$$\hat{m}_3 = \frac{\sum_{i=1}^n (X_i - \bar{X})^3}{n}$$

$$\hat{m}_4 = \frac{\sum_{i=1}^n (X_i - \bar{X})^4}{n}$$

Se define el tercer y cuarto momento estándar de una distribución como:

$$A_3 = \frac{m_3}{\sigma^3} = \frac{E\{(X - \mu)^3\}}{\left[E\{(X - \mu)^3\}\right]^{\frac{3}{2}}}$$

$$A_4 = \frac{m_4}{\sigma^4} = \frac{E\{(X - \mu)^4\}}{\left[E\{(X - \mu)^2\}\right]^2}$$

En el caso particular de la distribución normal  $N(\mu, \sigma)$  valen:  $\begin{cases} A_3 = 0 \\ A_4 = 3 \end{cases}$ . En general, si la

distribución es simétrica respecto de su media,  $\mu$ , el *coeficiente de simetría*  $A_3$  vale cero. El *coeficiente de kurtosis*  $A_4$  da idea del apuntamiento de la distribución, es decir del reparto del peso estadístico entre el centro y las colas de la distribución.

A partir de la muestra  $X_1, \dots, X_n$  pueden estimarse los coeficientes de simetría y kurtosis de la forma:

$$\hat{A}_3 = \frac{\hat{m}_3}{s^3}$$

$$\hat{A}_4 = \frac{\hat{m}_4}{s^4}$$

y comparar los valores obtenidos con los correspondientes a la distribución normal.

### 5.1.5- Ejercicio. Aplicación a un caso práctico

En cierto modelo de un aparcamiento público, los intervalos de tiempo entre llegadas sucesivas de vehículos son variables aleatorias continuas de entrada. Para poder simular el modelo es preciso identificar cómo están distribuidas estas variables aleatorias continuas.

Con este fin, se han recogido, durante 90 minutos, datos experimentales de los intervalos entre llegadas sucesivas de vehículos. Durante este periodo, han llegado 220 automóviles.

En la siguiente tabla se dan los 219 intervalos medidos entre estas llegadas, expresados en minutos y ordenados crecientemente. Se ha hecho coincidir el inicio de las observaciones con la llegada de un vehículo.

0.01	0.06	0.12	0.23	0.38	0.53	0.88
0.01	0.07	0.12	0.23	0.38	0.53	0.88
0.01	0.07	0.12	0.24	0.38	0.54	0.90
0.01	0.07	0.13	0.25	0.39	0.54	0.93
0.01	0.07	0.13	0.25	0.40	0.55	0.93
0.01	0.07	0.14	0.25	0.40	0.55	0.95
0.01	0.07	0.14	0.25	0.41	0.56	0.97
0.01	0.07	0.14	0.25	0.41	0.57	1.03
0.02	0.07	0.14	0.26	0.43	0.57	1.05
0.02	0.07	0.15	0.26	0.43	0.60	1.05
0.03	0.07	0.15	0.26	0.43	0.61	1.06
0.03	0.08	0.15	0.26	0.44	0.61	1.09
0.03	0.08	0.15	0.26	0.45	0.63	1.10
0.04	0.08	0.15	0.27	0.45	0.63	1.11
0.04	0.08	0.15	0.28	0.46	0.64	1.12
0.04	0.09	0.17	0.28	0.47	0.65	1.17
0.04	0.09	0.18	0.29	0.47	0.65	1.18
0.04	0.10	0.19	0.29	0.47	0.65	1.24
0.04	0.10	0.19	0.30	0.48	0.69	1.24
0.05	0.10	0.19	0.31	0.49	0.69	1.28
0.05	0.10	0.20	0.31	0.49	0.70	1.33
0.05	0.10	0.21	0.32	0.49	0.72	1.38
0.05	0.10	0.21	0.35	0.49	0.72	1.44
0.05	0.10	0.21	0.35	0.50	0.72	1.51
0.05	0.10	0.21	0.35	0.50	0.74	1.72
0.05	0.10	0.21	0.36	0.50	0.75	1.83
0.05	0.11	0.22	0.36	0.51	0.76	1.96
0.05	0.11	0.22	0.36	0.51	0.77	
0.05	0.11	0.22	0.37	0.51	0.79	
0.06	0.11	0.23	0.37	0.52	0.84	
0.06	0.11	0.23	0.38	0.52	0.86	
0.06	0.12	0.23	0.38	0.53	0.87	

219 intervalos entre llegadas ordenados crecientemente

Dibujar algunos histogramas de los datos experimentales. ¿Puede extraerse de ellos alguna conclusión?. Dibujar un gráfico de probabilidad de los datos con la distribución exponencial y otro con la distribución normal. ¿A que distribución se ajustan mejor los datos?.

El ejercicio continúa en el apartado 4.2.1.

## 5.2- Estimadores de máxima verosimilitud

Una vez que se ha seleccionado la familia de distribuciones a la que se va a realizar el ajuste, debemos estimar todos sus parámetros, para determinar la distribución de la que supuestamente hemos estado recogiendo muestras durante el proceso de medida. Empleamos los datos experimentales IID, tanto para decidir la familia de distribuciones, como para estimar los parámetros de la distribución.

Se llama *estimador* a una función numérica de los datos. Existen diferentes estimadores que dan el valor de un determinado parámetro de una distribución, en función de los datos experimentales. Existen, asimismo, muchos modos alternativos de obtener el valor de un estimador.

Nosotros trataremos solamente los *estimadores de máxima verosimilitud*, MLEs (maximum-likelihood estimators), si bien existen muchos otros, como los estimadores de mínimos cuadrados, los estimadores no sesgados, el método de los momentos, etc. El motivo de esta elección es que la idea del método es muy intuitiva y que es una técnica con propiedades muy buenas, a menudo no compartidas por los demás métodos de estimación.

Los fundamentos del método se comprenden mejor en el caso discreto. Supongamos que hemos seleccionado una distribución discreta para nuestros datos, con un parámetro desconocido  $\theta$ . Sea  $p_\theta(x)$  la probabilidad de esta distribución. Definimos la función verosimilitud  $L(\theta)$ , a partir de los datos experimentales IID,  $X_1, \dots, X_n$ , de la forma:

$$L(\theta) = p_\theta(X_1)p_\theta(X_2)\dots p_\theta(X_n)$$

con lo cual, dado que los datos experimentales son independientes, es igual a la probabilidad de obtener esos datos si  $\theta$  es el valor del parámetro desconocido. Entonces, el estimador de máxima verosimilitud (MLE) del parámetro desconocido  $\theta$ , que notaremos  $\hat{\theta}$ , se define como el valor de  $\theta$  que maximiza  $L(\theta)$ ; es decir,  $L(\theta) \leq L(\hat{\theta})$  para todos los posibles valores de  $\theta$ .

En el caso continuo, no puede darse una explicación tan intuitiva, ya que la probabilidad de que una variable aleatoria tome un valor determinado es siempre cero. En el caso continuo, si  $f_\theta(x)$  es la densidad de probabilidad de la distribución con un parámetro desconocido,  $\theta$ , la función de verosimilitud se define:

$$L(\theta) = f_\theta(X_1)f_\theta(X_2)\dots f_\theta(X_n)$$

El estimador de máxima verosimilitud,  $\hat{\theta}$ , del parámetro desconocido  $\theta$ , se define como el valor de  $\theta$  que maximiza  $L(\theta)$  para todos los valores permitidos de  $\theta$ .

Algunas de las propiedades de este método son:

- Para la mayoría de las distribuciones el estimador,  $\hat{\theta}$ , es único; es decir,  $L(\hat{\theta})$  es estrictamente mayor que  $L(\theta)$  para cualquier otro valor admisible de  $\theta$ . Generalmente los estimadores máximo-verosímiles pueden obtenerse mediante métodos de cálculo, ya que el máximo relativo de la función de verosimilitud obtenido diferenciando  $L(\theta)$  con respecto a  $\theta$  e igualando la derivada a cero es, generalmente, un máximo absoluto.

- Una propiedad deseable en una estimación es que esta "converja" hacia el valor verdadero del parámetro a medida que el tamaño de la muestra aumente. Como casi cualquier estimación razonable poseerá esta propiedad, en su lugar se impone a menudo una propiedad íntimamente relacionada con ésta, pero algo más restrictiva: que el estimador no este sesgado. Se dice que el estimador no esta sesgado cuando la media (o valor esperado) de la distribución asintótica de  $\hat{\theta}$  (cuando  $n \rightarrow \infty$ ) vale  $\theta$ , es decir, el parámetro que queremos estimar. Es decir, cuando  $n \rightarrow \infty$  la variable aleatoria  $\hat{\theta}$  tiene una distribución cuya media es el parámetro que se esta estimando,  $\theta$ .

- El estimador es invariante, es decir, si  $\phi = h(\theta)$ , donde h puede ser cualquier función, entonces  $\hat{\phi} = h(\hat{\theta})$ .

Consideremos el problema de estimar el parámetro  $\beta$  en la función de densidad exponencial

$$f(x; \beta) = \frac{1}{\beta} e^{-\frac{x}{\beta}}, x > 0, \beta > 0, \text{ si cinco observaciones de X dieron los valores:}$$

$$x_1 = 0.9, x_2 = 1.7, x_3 = 0.4, x_4 = 0.3, x_5 = 2.4$$

La función de verosimilitud en este caso es:

$$L = \frac{1}{\beta} e^{-\frac{x_1}{\beta}} \frac{1}{\beta} e^{-\frac{x_2}{\beta}} \dots \frac{1}{\beta} e^{-\frac{x_n}{\beta}} = \left( \frac{1}{\beta} \right)^n e^{-\frac{\sum_{i=1}^n x_i}{\beta}}$$

Puesto que el valor de  $\beta$  que maximiza L es el mismo que el que maximiza  $\log(L)$ , y como este último es más fácil de obtener, se calcula primero  $\log(L)$ :

$$\log(L) = -n \log(\beta) - \frac{1}{\beta} \sum_{i=1}^n x_i$$

Diferenciando respecto a  $\beta$  e igualando la derivada a cero, obtenemos la ecuación:

$$-\frac{n}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^n x_i = 0$$

cuya solución nos da un extremo (máximo, mínimo o punto de inflexión): da el estimador máximo-verosimil deseado de  $\beta$ :

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i}{n}$$

que, es igual a la media aritmética de las  $x_i$ . Examinando la segunda derivada particularizada

$$\text{para } \beta = \hat{\beta} = \frac{\sum_{i=1}^n x_i}{n}$$



$$\left[ \frac{n}{\beta^2} - 2 \frac{1}{\beta^3} \sum_{i:1}^n x_i \right]_{\beta=\hat{\beta}} = \frac{n}{\hat{\beta}^2} - 2 \frac{1}{\hat{\beta}^3} \sum_{i:1}^n x_i = -\frac{n}{\hat{\beta}^2} < 0 \quad \text{para todo } \hat{\beta}$$

vemos que el extremo es un máximo.

La estimación máximo-verosimil del parámetro de la distribución exponencial que se ajusta a las medidas es:

$$\hat{\beta} = \frac{0.9 + 1.7 + 0.4 + 0.3 + 2.4}{5} = 1.14$$

Como ilustración de un caso para el cual los métodos de cálculo son inadecuados, consideremos el problema de hallar la estimación máximo-verosimil del parámetro  $\theta$  para la densidad de probabilidad uniforme  $f(x; \theta) = \frac{1}{\theta}$  con  $0 \leq x \leq \theta$ , basada en los valores muestrales  $x_1, \dots, x_n$ .

En este caso:

$$L(\theta) = \prod_{i:1}^n f(x_i; \theta) = \left( \frac{1}{\theta} \right)^n$$

Esta función será máxima eligiendo  $\theta$  tan pequeña como sea posible, bajo la restricción  $0 \leq x_i \leq \theta$ ,  $i = 1, \dots, n$ . El menor valor de  $\theta$  que satisface estas desigualdades es el mayor valor de las  $x_i$ . Así pues, la estimación máximo-verosimil de  $\theta$  viene dada por:  $\hat{\theta} = \max\{x_1, \dots, x_n\}$

Hasta ahora hemos tratado distribuciones con un solo parámetro desconocido. Si la distribución tiene varios parámetros pueden definirse MLEs de forma análoga. Por ejemplo, para la distribución gamma la función de máxima-verosimilitud se define:

$$L(\alpha, \beta) = \frac{x_1^{\alpha-1} e^{-\frac{x_1}{\beta}}}{\beta^\alpha \Gamma(\alpha)} \frac{x_2^{\alpha-1} e^{-\frac{x_2}{\beta}}}{\beta^\alpha \Gamma(\alpha)} \cdots \frac{x_n^{\alpha-1} e^{-\frac{x_n}{\beta}}}{\beta^\alpha \Gamma(\alpha)}$$

Los estimadores de máxima verosimilitud  $\hat{\alpha}$  y  $\hat{\beta}$  son los valores de  $\alpha$  y  $\beta$  que hacen que  $L(\alpha, \beta)$  sea máximo. Para ello, debe calcularse  $\alpha$  y  $\beta$  que resuelve el sistema de ecuaciones:

$$\begin{cases} \frac{\partial}{\partial \alpha} \ln(L(\alpha, \beta)) = 0 \\ \frac{\partial}{\partial \beta} \ln(L(\alpha, \beta)) = 0 \end{cases}$$

### 5.2.1- Ejercicio. (Continuación)

Estimar el valor del parámetro de la distribución exponencial que mejor se ajusta, según el criterio de máxima verosimilitud, a los datos experimentales dados en el ejercicio 2. Continúa en la sección 4.3.1.

### 5.3- Medida de la bondad de ajuste. Test $\chi^2$ (chi-cuadrado)

Una vez que se han estimado los parámetros de la distribución, a partir de los datos experimentales, debemos determinar en qué medida los datos experimentales responden a la distribución ajustada,  $\hat{F}$ . Se trata de responder a la pregunta: ¿cabría haber obtenido los datos experimentales muestreando la distribución ajustada?. Los "test de ajuste" pretenden comprobar la hipótesis:  $H_0$ : Las variables aleatorias  $X_1, \dots, X_n$  IID tienen distribución  $\hat{F}$ .

En efecto,  $H_0$  nunca será literalmente cierta: probablemente nunca obtendremos la distribución exacta de las observaciones  $X_1, \dots, X_n$ . Los test de ajuste deben interpretarse como un método sistemático de detectar grandes discrepancias entre la distribución ajustada y los datos experimentales. De hecho, la mayoría de los test no son demasiado potentes: cuando se dispone de pocas observaciones son poco sensibles a discrepancias entre los datos y la distribución ajustada y, sin embargo, cuando se dispone de muchas observaciones una pequeña discrepancia puede hacer que se rechace el ajuste.

La bibliografía sobre test de ajuste es muy extensa. Nos limitaremos a describir uno de ellos, el test chi-cuadrado, si bien no debemos perder de vista que existen muchos otros test para distribuciones genéricas (test de Kolmogorov-Smirnov, de Anderson-Darling, de Cramer-von Mises, etc...) y test diseñados para distribuciones específicas determinadas.

El test  $\chi^2$  (chi-cuadrado) puede considerarse como un método formal de comparar el histograma o diagrama de barras de los datos experimentales, con la función densidad de probabilidad o probabilidad de la distribución ajustada.

Debe dividirse el rango de la distribución ajustada en  $k$  intervalos adyacentes  $[a_0, a_1), [a_1, a_2), \dots, [a_{k-1}, a_k)$ , donde puede suceder que  $a_0 = -\infty$ , en cuyo caso el primer intervalo será  $(-\infty, a_1)$ , o que  $a_k = \infty$ , o ambas cosas.

Se define  $N_j$  como el número de los  $X_i$ 's en el intervalo  $j$ -ésimo,  $[a_{j-1}, a_j)$ , para  $j:1, \dots, k$ . Se satisfará  $\sum_{j:1}^k N_j = n$ . A continuación se calcula la proporción de los  $X_i$ 's,  $p_j$ , que caerían en el intervalo  $j$ -ésimo si el muestreo si hiciera de la distribución ajustada.

En el caso continuo, notando  $\hat{f}(x)$  la densidad de probabilidad de la distribución ajustada:

$$p_j = \int_{a_{j-1}}^{a_j} \hat{f}(x) dx$$

En el caso discreto, notando  $\hat{p}$  la probabilidad de la distribución ajustada:

$$p_j = \sum_{\{i: a_{j-1} \leq x_i < a_j\}} \hat{p}(x_i)$$

Para realizar el contraste debe calcularse el valor de:

$$\chi^2 = \sum_{j:1}^k \frac{(N_j - np_j)^2}{np_j}$$

Dado que  $np_j$  es el número de datos que caerían en el intervalo  $j$ -ésimo si el ajuste fuera perfecto, cabe esperar que cuanto mejor sea el ajuste, menor sea  $\chi^2$ . Con lo cual, rechazaremos la hipótesis  $H_0$  si  $\chi^2$  es demasiado grande.

La forma precisa del test depende de si hemos estimado alguno de los parámetros de la distribución por ajuste a partir de los datos experimentales (que es nuestro caso) o si, por el contrario, es un dato conocido. El número de grados de libertad es igual a:

$$k - 1 - \left\{ \begin{array}{l} \text{Numero de parametros de la distribucion} \\ \text{estimados de los datos experimentales} \end{array} \right\}$$

Si se ha estimado la media y la varianza de la distribución a partir de la muestra, debe emplearse  $\chi^2_{k-3,1-\alpha}$  para contrastar la hipótesis:

$$\left\{ \begin{array}{ll} \chi^2 > \chi^2_{k-3,1-\alpha} & \text{se rechaza } H_0 \\ \chi^2 \leq \chi^2_{k-3,1-\alpha} & \text{no hay evidencia para rechazar } H_0 \end{array} \right.$$

El aspecto más problemático del test es la elección de los intervalos, ya que no existe un método genérico óptimo que garantice buenos resultados para cualquier distribución ajustada y cualquier tamaño de muestra. Existen, sin embargo, algunas recomendaciones que son generalmente aceptadas:

- Se recomienda escoger los intervalos de modo que  $p_1 = p_2 = \dots = p_k$  o, al menos, de modo que sean aproximadamente iguales. En el caso discreto, generalmente no será posible hacer las  $p_j$ 's iguales; en el caso continuo, conviene seguir este consejo, dado que la distribución de probabilidad acumulada ajustada debe ser invertida.

Una razón de recomendar que las  $p_j$ 's sean iguales es que esto hace que el test no este sesgado. El test no esta sesgado cuando es más probable rechazar la hipótesis  $H_0$  cuando es falsa que cuando es verdadera, con lo cual, un test sesgado no es útil.

- Se recomienda escoger los intervalos de modo que los valores  $np_j$  no sean demasiado pequeños. Una regla práctica ampliamente empleada es escoger  $np_j \geq 5$  para todo (o casi todo)  $j$ . La razón de esta recomendación es que la coincidencia entre la distribución verdadera  $\chi^2$  (para  $n$  fijo y finito) y su distribución asintótica (cuando  $n$  tiende a infinito) chi-cuadrado (usada para obtener el valor crítico del test) es mejor si los valores  $np_j$  no son demasiado pequeños.

Como vemos, existen razones para escoger los intervalos de modo que los  $p_j$ 's sean aproximadamente iguales entre ellos y de modo que  $np_j \geq 5$  para toda  $j$  excepto, a lo sumo, para una o dos de ellas. Sin embargo, estas recomendaciones no proporcionan una respuesta completa al problema de la elección del intervalo; en particular, no hemos dicho nada acerca de cómo decidir en número de intervalos  $k$ .

La elección de  $k$  afecta principalmente a la potencia del test, si bien el valor de  $k$  que conduce a la máxima potencia depende de la forma de la distribución. En general, suele aconsejarse no escoger  $k$  mayor de 30 ó 40. Por desgracia, no existe una respuesta simple y eficaz a esta pregunta y este es uno de los principales inconvenientes del test chi-cuadrado. En algunas ocasiones, pueden obtenerse conclusiones completamente diferentes, a partir del mismo conjunto de observaciones, dependiendo de la elección de los intervalos. Pese a ello, el test chi-cuadrado es ampliamente utilizado y es aplicable a cualquier tipo de distribución ajustada.

**5.3.1- Ejercicio. (continuación).** Emplear el test chi-cuadrado para comparar los 219 intervalos entre llegadas, con la distribución exponencial ajustada. Escoger  $k=20$  y  $\alpha = 0.10$ .

## 5.4- Selección de una distribución en ausencia de datos

En algunos estudios de simulación, no es posible recoger datos de las variables aleatorias de interés, con lo cual no pueden emplearse las técnicas anteriormente discutidas para seleccionar la distribución. En esta sección discutiremos tres aproximaciones ampliamente empleadas para seleccionar distribuciones en ausencia de datos: suponer que  $X$  tiene una función densidad de probabilidad triangular, uniforme o beta.

Supongamos que la variable aleatoria continua de interés,  $X$ , es el tiempo empleado en realizar una tarea (tiempo hasta que se produce el fallo de un componente, tiempo de fabricación de un producto,...). El primer paso es estimar el intervalo  $[a,b]$  en el cual se encuentra  $X$  con probabilidad 1, es decir,  $P\{X < a \text{ o } X > b\} \approx 0$ . Los extremos  $a, b$  del intervalo son las estimaciones más optimista y pesimista, respectivamente, del tiempo necesario para realizar la tarea. Una vez estimado  $[a,b]$ , debe definirse en el intervalo una función densidad de probabilidad que sea representativa de  $X$ .

Para aproximar la función densidad de probabilidad de  $X$  por una triangular, es preciso estimar el tiempo más probable de realización de la tarea. Este valor más probable,  $c$ , es el parámetro de forma de la distribución.

Si puede considerarse que  $X$  está igualmente distribuida en  $[a,b]$ , puede aproximarse la función densidad de probabilidad de  $X$  por la uniforme  $U(a,b)$ .

Otra aproximación, más flexible, consiste en suponer que  $X$  tiene en  $[a,b]$  una función densidad de probabilidad de parámetros de forma  $\alpha_1, \alpha_2$ . Para ello es útil conocer algunas propiedades de las distribuciones beta:

- Una variable aleatoria beta de rango  $[0,1]$  puede reescalarse y desplazarse para obtener una variable aleatoria beta de rango  $[a,b]$  con la misma forma mediante la transformación  $a + (b - a)X$ .
- Si  $X$  es una variable aleatoria beta de rango  $[0,1]$ ,  $beta(\alpha_1, \alpha_2)$ , entonces la variable aleatoria  $(1 - X)$  es una variable aleatoria  $beta(\alpha_2, \alpha_1)$  de rango  $[0,1]$ .
- La función densidad de probabilidad, con rango  $[0,1]$ , es simétrica respecto a  $x = \frac{1}{2}$  si y sólo si  $\alpha_1 = \alpha_2$ . La media y el modo (valor de  $X$  para el cual la densidad de probabilidad es máxima) son iguales si y sólo si  $\alpha_1 = \alpha_2$ .
- $U(0,1)$  y  $beta(1,1)$  son la misma distribución.

- beta(1,2) y triang(0,1,0) son la misma distribución. beta(2,1) y triang(0,1,1) son la misma distribución.

- Si  $\alpha_1 > \alpha_2$ , con rango  $[0,1]$ ,  $P\{X \leq 0.5\} < P\{X \geq 0.5\}$ . Por el contrario, si  $\alpha_1 < \alpha_2$ ,  $P\{X \leq 0.5\} > P\{X \geq 0.5\}$ .

- El modo de la densidad de distribución con rango  $[0,1]$  es:

$$\left\{ \begin{array}{ll} \frac{\alpha_1 - 1}{\alpha_1 + \alpha_2 - 2} & \text{si } \alpha_1 > 1, \alpha_2 > 1 \\ 0 \text{ y } 1 & \text{si } \alpha_1 < 1, \alpha_2 < 1 \\ 0 & \text{si } \alpha_1 < 1, \alpha_2 \geq 1 \text{ o si } \alpha_1 = 1, \alpha_2 > 1 \\ 1 & \text{si } \alpha_1 \geq 1, \alpha_2 < 1 \text{ o si } \alpha_1 > 1, \alpha_2 = 1 \\ \text{no es unico} & \text{si } \alpha_1 = \alpha_2 = 1 \end{array} \right.$$

## 5.5- Modelos de procesos de llegada. Procesos de Poisson

### 5.5.1- Procesos de Poisson estacionarios

En muchas simulaciones necesitamos generar una secuencia de puntos aleatorios en el tiempo,  $0 = t_0 \leq t_1 \leq t_2 \leq \dots$  de modo que el evento  $i$ -ésimo de algún tipo ocurre en el instante  $t_i$  y que la distribución de los eventos temporales  $\{t_i\}$  es de una forma determinada. Sea  $N(t) = \max\{i: t_i \leq t\}$  el número de eventos que suceden hasta el instante  $t$ , con  $t \geq 0$ . En lo sucesivo, llamaremos al proceso estocástico  $\{N(t), t \geq 0\}$  *proceso de llegada*, ya que, para nuestros propósitos, los eventos de interés son llegadas de usuarios a algún tipo de servicio. Asimismo, llamaremos a  $A_i = t_i - t_{i-1}$   $\{i: 1, 2, \dots\}$  *intervalo entre llegadas* entre el usuario  $(i-1)$ -ésimo y el  $i$ -ésimo.

Los *procesos de Poisson* son procesos de llegada en los cuales las variables aleatorias  $A_i$ 's son IID exponenciales. La llegada de clientes a un sistema de colas suele modelarse como un proceso de Poisson. Un proceso estocástico  $\{N(t), t \geq 0\}$  se llama proceso de Poisson (estacionario) si:

- 1.- Los usuarios llegan uno a uno.
- 2.-  $N(t+s) - N(t)$ , que es el número de llegadas en el intervalo  $(t, t+s]$ , es independiente de  $\{N(u), 0 \leq u \leq t\}$
- 3.- La distribución de  $N(t+s) - N(t)$  es independiente de  $t$  para todo  $t$ ,  $s \geq 0$ . El proceso de Poisson es estacionario cuando cumple esta condición.

Las dos primeras propiedades son características de muchos procesos de llegada. La propiedad 1 impide que los usuarios lleguen al sistema en grupos. La propiedad 2 dice que el número de llegadas en el intervalo  $(t, t+s]$  es independiente del número de llegadas en el intervalo  $[0, t]$  y de los instantes en que se produjeron esas llegadas. Estas propiedad se vería violada si, por ejemplo, un número grande de llegadas en  $[0, t]$  hiciera que los usuarios llegados en  $(t, t+s]$  encontraran el sistema congestionado y decidieran "volver otro día".

La propiedad 3 no es satisfecha por la mayoría de los sistemas reales, ya que supone que la frecuencia de llegada es independiente de la hora del día. Sin embargo, si el periodo de interés del sistema es relativamente corto, de modo que pueda considerarse en él la frecuencia de llegadas constante, el proceso durante ese intervalo puede modelarse como uno de Poisson.

El siguiente teorema explica el por qué se llaman procesos de Poisson:

Si  $\{N(t), t \geq 0\}$  es un proceso de Poisson, entonces el número de llegadas en cualquier intervalo de longitud  $s$  es una variable aleatoria de Poisson con parámetro  $\lambda s$ , donde  $\lambda$  es un real positivo. Esto es:

$$P\{N(t+s) - N(t) = k\} = \frac{e^{-\lambda s} (\lambda s)^k}{k!} \quad \text{para } k = 0, 1, 2, \dots \text{ y } t, s \geq 0$$

donde:  $E(N(s)) = \lambda s$ . En particular,  $E(N(1)) = \lambda$ , donde vemos que  $\lambda$  es el número esperado de llegadas en cualquier intervalo de longitud uno.  $\lambda$  se llama *frecuencia de llegadas* del proceso.

El siguiente teorema establece que los intervalos entre llegadas en un proceso de Poisson son variables aleatorias IID exponenciales:

Si  $\{N(t), t \geq 0\}$  es un proceso de Poisson con frecuencia  $\lambda$ , entonces sus correspondientes intervalos entre llegadas  $A_1, A_2, \dots$  son variables aleatorias exponenciales IID con parámetro  $\frac{1}{\lambda}$ , es decir,  $\text{expo}(\frac{1}{\lambda})$ .

El inverso de este teorema también es cierto: si los intervalos entre llegadas  $A_1, A_2, \dots$  de un proceso de llegadas  $\{N(t), t \geq 0\}$  son variables aleatorias IID exponenciales con parámetro  $\frac{1}{\lambda}$ , entonces  $\{N(t), t \geq 0\}$  es un proceso de Poisson con frecuencia  $\lambda$ .

### 5.5.2- Procesos de Poisson no estacionarios

Sea  $\lambda(t)$  la frecuencia de llegada de usuarios en el instante  $t$ . Si los usuarios llegan al sistema de acuerdo con un proceso de Poisson con frecuencia  $\lambda$ , entonces  $\lambda(t) = \lambda$  para todo  $t \geq 0$ . Sin embargo, en la mayoría de los procesos reales,  $\lambda(t)$  no es constante, con lo cual, los intervalos entre llegadas  $A_1, A_2, \dots$  no están idénticamente distribuidos: no es correcto ajustar una única distribución a todos los  $A_j$ 's usando las técnicas anteriormente descritas.

Un proceso estocástico  $\{N(t), t \geq 0\}$  se dice un *proceso de Poisson no estacionario* si:

- 1.- Los usuarios llegan uno a uno.
- 2.-  $N(t+s) - N(t)$ , que es el número de llegadas en el intervalo  $(t, t+s]$ , es independiente de  $\{N(u), 0 \leq u \leq t\}$

Como vemos, en un proceso de Poisson no estacionario, la frecuencia de llegadas es dependiente del tiempo. Cuando no lo es, se habla de proceso de Poisson no estacionario o, simplemente, de proceso de Poisson.

Se define la *función expectación*  $a(t)$  del proceso de Poisson no estacionario, como  $a(t) = E(N(t))$ , para todo  $t \geq 0$ . Si  $a(t)$  es diferenciable para un valor determinado de  $t$ , definimos la frecuencia de llegadas como:

$$\lambda(t) = \frac{d}{dt} a(t)$$

Intuitivamente,  $\lambda(t)$  será mayor en aquellos intervalos en los cuales el número esperado de llegadas sea mayor.

El siguiente teorema establece que el número de llegadas en el intervalo  $(t, t+s]$ , para un proceso de Poisson no estacionario, es una variable aleatoria de Poisson cuyo parámetro depende de  $t$  y de  $s$ .

Si  $\{N(t), t \geq 0\}$  es un proceso de Poisson no estacionario, con una función expectación  $a(t)$  continua, entonces:

$$P\{N(t+s) - N(t) = k\} = \frac{e^{-b(t,s)} (b(t,s))^k}{k!} \quad \text{para } k = 0, 1, 2, \dots \text{ y } t, s \geq 0$$

donde:  $b(t,s) = a(t+s) - a(t) = \int_t^{t+s} \lambda(y) dy$ , la última igualdad es cierta si  $\frac{da(t)}{dt}$  esta acotada en  $[t, t+s]$  y si  $\frac{da(t)}{dt}$  existe y es continua en el intervalo  $[t, t+s]$ , excepto, a lo sumo, un número finito de puntos.

En el siguiente ejemplo se expone un método de estimar  $\lambda(t)$  (o  $a(t)$ ) a partir de un conjunto de observaciones del proceso de llegadas de interés.

**Ejemplo.** A fin de desarrollar un modelo de simulación de una tienda de fotocopias, se recogen datos sobre los instantes de llegada de clientes entre las 11 de la mañana y la 1 de la tarde durante 8 días. La observación de las características de llegada de los clientes permiten suponer que las propiedades 1 y 2 de un proceso de Poisson se satisfacen y que  $\lambda(t)$  varía en el intervalo de 2 horas. Para obtener una estimación de  $\lambda(t)$ , se divide el intervalo de dos horas en los siguientes 12 subintervalos de 10 minutos de duración cada uno:

$$[11:00, 11:10), [11:10, 11:20), \dots, [12:40, 12:50), [12:50, 1:00]$$

Se calcula, para cada día, el número de llegadas en cada uno de estos subintervalos.

Se calcula, para cada subintervalo, el número medio de llegadas en los 8 días en que se han realizado observaciones. Estos 12 promedios son estimaciones del número de llegadas esperado en los correspondientes subintervalos.

Para cada subintervalo, se divide el promedio de llegadas que se produce en él, por la longitud del subintervalo (10 minutos), para obtener una estimación de la frecuencia de llegada en cada subintervalo,  $\hat{\lambda}(t)$ , medida en clientes por minuto.

Cabe preguntarse por qué se ha decidido que la longitud de los subintervalos sea 10 minutos. La elección, efectivamente, ha sido arbitraria. El problema de la elección del subintervalo en este caso es similar al planteado al dibujar un histograma.



### 5.5.3- Procesos de Poisson compuestos

En algunos sistemas reales, los clientes pueden llegar al sistema en grupos, de modo que la propiedad 1 de los procesos de Poisson no se satisface. Por ejemplo, los clientes suelen entrar en una cafetería en grupos. Consideremos cómo pueden modelarse este tipo de *procesos de llegada en grupos*.

Sea  $N(t)$  el número de grupos de usuarios que han llegado al sistema en el instante  $t$ . Aplicando las técnicas discutidas anteriormente a los intervalos entre llegadas sucesivas de grupos, puede desarrollarse un modelo para el proceso  $\{N(t), t \geq 0\}$ . Por ejemplo, si los intervalos entre las llegadas consecutivas de grupos de usuarios son variables aleatorias IID exponenciales,  $\{N(t), t \geq 0\}$  puede modelarse como un proceso de Poisson. A continuación, puede ajustarse una distribución discreta al tamaño de los sucesivos grupos de usuarios.

Si  $X(t)$  es el número total de usuarios que han llegado hasta el instante  $t$ , y si  $B_i$  es el número de usuarios en el grupo  $i$ -ésimo, entonces  $X(t)$  viene dado por:

$$X(t) = \sum_{i:1}^{N(t)} B_i \quad \text{para } t \geq 0$$

Si las variables aleatorias  $B_i$  son IID y además son independientes de  $\{N(t), t \geq 0\}$ , y si  $\{N(t), t \geq 0\}$  es un proceso de Poisson, entonces el proceso estocástico  $\{X(t), t \geq 0\}$  se dice un *proceso de Poisson compuesto*.

Supongamos que observamos un proceso de Poisson durante un intervalo fijo de tiempo  $[0, T]$ , donde  $T$  es una constante determinada antes de iniciar las observaciones. Sea  $n$  el número de eventos que observamos en el intervalo  $[0, T]$  y sea  $t_i$  el instante en que sucede el evento  $i$ -ésimo ( $i:1, 2, \dots, n$ ). Se cumple  $0 \leq t_1 \leq t_2 \leq \dots \leq t_n \leq T$ . Si  $t_n < T$ , no suceden eventos en el intervalo  $(t_n, T]$ . Veamos el modo en que esta relacionada la distribución de  $t_1, t_2, \dots, t_n$  con la distribución  $U(0, T)$ .

Supongamos que  $Y_1, Y_2, \dots, Y_n$  (la misma  $n$  que antes) son variables aleatorias IID con distribución  $U(0, T)$ . Sea  $Y_{(i)}$  la  $i$ -ésima más pequeña de las  $Y_j$ 's, es decir,  $Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(n)}$ . Entonces, una propiedad de los procesos de Poisson es que  $t_1, t_2, \dots, t_n$  tienen la misma distribución que  $Y_{(1)}, Y_{(2)}, \dots, Y_{(n)}$ . Un modo de interpretar esta propiedad es que a la vista de los valores  $t_1, t_2, \dots, t_n$ , no podríamos diferenciar si  $t_i$  el instante en que sucede el evento  $i$ -ésimo de una secuencia de eventos o si son muestras de  $n$  variables aleatorias IID de distribución  $U(0, T)$  ordenadas en orden creciente. Asimismo, considerando  $t_1, t_2, \dots, t_n$  como variables aleatorias desordenadas, son IID con una distribución  $U(0, T)$ .

Esta propiedad indica que comprobar la hipótesis de que una secuencia observada de eventos es generada por un proceso de Poisson es equivalente a comprobar que los instantes de los eventos,  $t_1, t_2, \dots, t_n$ , son variables aleatorias  $U(0, T)$  IID.